DOE Resources and Facilities for Biological Discovery in the 21st Century Realizing the Potential Enabled by the Genome Projects

A working paper presented to the BERAC, 4/25/02

Summary

The genomic revolution, driven by the Human and other Genome Projects, has opened a new era in the science and applications of biology an era of systems biology. We can now bring together the concepts, and technologies of the biological, physical, and computing sciences to enable a comprehensive and fundamental understanding of life. This revolution will translate into many applications and deliver enormous benefit to the nation and the world. The DOE should now create unique, high throughput research facilities and resources to translate the new biology, embodied in the Genomes to Life (GTL) program, into a reality for the nation. The BER program of DOE, having played a critical, catalytic role in bringing about the genomic revolution, is now poised to make equally seminal contributions to this next, transforming phase. These next steps are where the major payoff to the DOE missions in energy, climate, security, and the environment and to the nation resides.

The bold vision of the DOE BER Genomes to Life Program is designed to build on the major accomplishments of the past decade and move from this vision to reality to a new and comprehensive systems approach from which we will understand the functioning of cells and organisms and their interactions with their environments. Since the science has changed so profoundly, to accomplish these challenging goals in a timely and cost effective fashion, new facilities and new scientific resources are needed. The facilities can be separate new ones or be attached to existing facilities in the DOE system or in academic institutions. Examples include protein production facilities, data centers affiliated with the major computing facilities, centers to support computational biology, new or expanded facilities to provide capacity for an expanded user community in big instruments like Mass Spectrometry for proteomics, NMR, X-Ray, electron microscopes, lasers, and the like. The key is to understand the longer range goals of GTL and the extent to which existing research infrastructure must be expanded, redirected and /or complemented with a new set of core capabilities and technologies to achieve these goals. This document argues for and outlines a new plan that will provide the needed facilities and scientific infrastructure to support the BER’s Genomes to Life Program. We call for an immediate planning process including action by BERAC and ASCAC and workshops for evaluation and prioritorization of needs and prospective new technological advances.

The Genomic Revolution

Genome data provide a fundamental, new starting point for understanding Life’s processes because the genome contains the information necessary to create and sustain complex living systems. The intricacy of the interactions of the subsystems of a living cell is becoming increasingly evident. It is clear, for example, that the major elements of molecular function, proteins, rarely work alone, but rather act as elements of multi-molecular machines and pathways, in closely coupled physical networks. What the genome sequence has provided is the full list of the components and the basic regulatory information for the structure and function of the panoply of molecular machines active in the cell. This new knowledge, in combination with the advent of new technologies, has opened the door to a strategy based on comprehensive and high throughput experimental approaches to understanding the presence, identity, number,location, and function of all the working elements of a living system and the processes by which they are synthesized and controlled. At the heart of GTL is a new systems approach to biology enabled and inspired by the successes of the genome projects. The Genome projects have yielded four fundamental lessons: the value of high throughput biology, economies of scale, the power of discovery-based research for hypothesis generation, and the enormous power of open data with rapid availability to the entire scientific community.

The new tools of genomics and the new technologies for probing cellular functions now enable discovery of the sophisticated ways that evolution has solved engineering and information handling problems. Now we can take advantage of the three billion of years of selection that nature has sustained. The success of structural biology programs at large synchrotron, facilities in the past decade clearly demonstrates of the value of creating high technology, large-scale resource centers. The BER program of DOE has played a critical, catalytic role in bringing about the genomic revolution. The genome project has its origins in BER (1987), the microbial genome project originated in BER, and BER is now poised to make equally seminal contributions to this next, transforming phase. These next steps are where the major payoff to the DOE mission and to the nation resides.

Our ability to capitalize on this information depends on the confluence of several specific advances, including advanced experimental tools such as mass spectrometry, x-rays, neutrons, and NMR to study structure and function, tools to study biological function (micro arrays, protein chips etc), high performance computing and simulations, and our nascent ability to integrate (both physically and conceptually) into a real systems approach. The integration of this knowledge and all of these capabilities and their availability to all scientists is enabled by the revolution in information and communications technologies and science a central nervous system for any conceivable facilities infrastructure.

DOE can create the needed interfaces between the requirements and aspirations of this broad vision and its extensive resources in biophysical, physical and computing sciences and engineering. Realizing this vision implies the imperative to marry experimental data analysis and experiment design with computation including a new generation of modeling and simulation. To truly understand the ways in which genetic information is translated into the functions of living systems, the infrastructure for intelligent experiment design and the development of modeling and simulation must be closely coupled -- each process informing the other.

Genomes to Life

To realize the full potential that the genome projects have provided - tounderstand how the parts encoded by the genome work together, the past research paradigm of studying isolated single-biological processes (one gene at a time, for example) is being transformed into a systems level paradigm for research. This new paradigm rests on three critical foundations;

1. Full genomic sequences of organisms and their evolutionary relatives, and a variety of information about molecular and cellular structure,

2. New technologies and techniques with their origins in physical, chemical and engineering sciences as well as the life sciences, to allow rapid, system-wide measurement on living systems, at the molecular level; and

3. Information and computing advances that provide scientists ready access to comprehensive information and the tools to incorporate that information into models to probe the processes and phenomena of living systems, test hypotheses and ideas, and inspire and inform new forms of experimental inquiry. The only robust approach to grapple with the complexities of living systems is through the temporal, spatial, and functional flexibility of computational simulation and modeling.

The Genomes to Life initiative seeks to bring these three elements together to create an environment to realistically and expeditiously pursue a working understanding of the many living systems that are important to the missions of DOE, particularly for payoffs for Energy, Climate, Security, and Environment. An integration of the three critical foundations will thus form a vital part of the new core of information and experimental capabilities for biological science and technology for all DOE programs, and be a critical enabler of the national effort in 21st Century Biology. The many needs and opportunities of the full DOE Science and Technology portfolio and the many benefits to other national programs requires that BER establish a robust and comprehensive resource base to support the growing needs of the scientists working on these complex problems. GTL represents both the motivation for and the nucleus of the new resources and facilities program.

A Call for New Resources

We argue here that a key part of this new enterprise is to make available the full suite of available technologies for individual researchers, as well as access the vast new information resources that will rapidly evolve. Achieving the great promise before us will require a sophisticated management philosophy that brings the life, physical, and computing disciplines together with a coherent set of goals to create this new environment. This enterprise must harness the unique powers and resources of the national laboratories, academia, and industry in new ways if the promise is to become a reality.

As the genome projects have shown, a comprehensive and high throughput approach to biology cannot only provide a revolutionary model and approach to understanding life, but also requires new institutional models and approaches that go far beyond the historical single investigator model that has been so important in the past. Key to success of GTL is genome-scale collection, analysis, dissemination and modeling of data. Just as with the HGP and the community generation of DNA sequence, a key to the success of the GTL will the generation, for the research community, of genome-scale data, and data management and analysis tools and capabilities for the biological "outputs" of a genome - biological function that results from genetic regulation, molecular machines, higher order structure and function of cells organisms and microbial communities. We argue that to make new capabilities widely available, and to use them effectively can best be done by establishing new Facilities and Resource Centers. These centers must serve the community of national lab, academic and industrial research users. Creating this new capability will require a concerted strategy that speaks to a wide range of institutional needs and requirements for DOE, the National Laboratories, Academia, Industry, and the other Federal Agencies.

The Rationale and Benefits for Facilities and Resource Centers:

DOE s GTL Facilities and Resource Centers should be developed and operated for a wide variety of specific purposes which include the following:

· to enable systems biology research to take full advantage of the existing national facilities;

· to assemble and facilitate new capabilities for high throughput systems biology;

· to create effective centers for biological computing and information management;

· to advance technology development to enable the implementation of GTL;

· to facilitate the application of GTL science and technologies to specific areas within DOE s science and technology portfolio especially specific DOE mission areas; and

· to provide resources and user facilities for scientists throughout the National Laboratories, Academia, other federal agencies, and industry.

Bringing together resources and people to economically create such Facilities and Resource Centers can have many benefits.

1. They will enable new science. Providing scientists with the ability to open new avenues of inquiry will fundamentally change the course biology in the coming decades. New kinds of questions can be asked and answered. This will attract the very best talent to the field.

2. They will stimulate multidisciplinary technology development. Many technologies can only be developed in an environment of deep and broad technical and engineering resources.

3. They could be the only way to do some things of great significance to biological discovery -- the synchrotrons sources represent a striking example of this phenomenon.

4. There is an economy and efficiency of scale for some important

resources as has been evident in genome sequencing centers like the DOE’s Joint Genome Institute.

5. They can bring together advanced and diverse technologies and programs in an integrated computing and information environment and thus provide a comprehensive infrastructure if multiple capabilities and resources are integrated effectively, much more effective research can result.

6. They can be the point of focus of whole new communities of scientists at the interfaces of disciplines.

7. They can break down the walls between institutions, providing a new venue for science that transcends National Laboratory/Academic/Industrial boundaries.

The development of major facilities has been a great success story for DOE, and has led to many revolutionary avenues of research and discovery for the nation. As biology progresses, and as the systems-wide and high throughput approach begun in the genome project becomes more prevalent, resources to enable this new science will become more important and BER must aspire to provide them. These new resources will include distributed networks of resources, enhancement of existing user facilities, and new stand-alone facilities for new purposes.

Management and Implementation: To assure excellence in conception and execution in the design and establishment of this suite of capabilities, an open and peer reviewed competitive process involving the broad user community is essential. The identification of new concepts, planning, and facilities development should first be carried out by a variety of mechanisms including workshops and BERAC and ASCAC panels and sub-committees convened for this purpose. There are numerous BES facilities which will also be critical for this program. Similar mechanisms, working with BESAC, will be important to exploit the links between the BER and BES programs and for facilities planning.

The creation of this new generation of facilities offers an opportunity to leverage the best in the individual investigator tradition of the life sciences and the tradition of creating sophisticated and technologically advanced facilities and computing infrastructure from the physical and computing sciences. In today s world of rapid communications and travel, an environment based on creativity and the entrepreneurial spirit with ready availability of the most advanced technologies and concepts should be attainable given appropriate management focus and funding. Some new facilities might be designated as pilot projects and graduated to full facilities when they pass an evaluative phase. It is important to determine whether there is a real advantage to centralizing a particular capability, and whether the initial location and team can deliver an effective resource to the community.

Facilities Requirements:

A. Existing Facilities:

As DOE is the agency primarily responsible for creating and operating major scientific user facilities, there is a robust assortment of facilities that are currently used by a very large multidisciplinary community. A key priority for GTL is the establishment of support, production, and other resources to assure access and optimal use of these existing facilities for GTL related research and applications.

In the case of the four synchrotrons (ALS, APS, NSLS, and SSRL), partnership with NIH and other non-federal partners has created a very effective and large set of instruments that serve the Nation s needs for doing frontier work in structural biology. Strong programs in structural genomics are being developed at all of these 4 facilities, primarily with funding from NIH. It can be well argued that the facilities nucleate the formation of strong, innovative scientific programs, a phenomenon we would hope to duplicate in other, new areas. The sensible approach for this area would be to enhance and make available existing capabilities to support the structural biology aspects of GTL.

In addition to the synchrotrons, existing DOE capabilities for biology include:

· Significant sequencing capability at the Joint Genome Institute, annotation and sequence finishing at ORNL, LANL, Stanford Medical School and the Comprehensive Microbial Data Base Center at TIGR;

· Forefront NMR capabilities and supporting isotopic labeling capabilities (PNNL and others);

· Forefront Mass Spectroscopy capabilities for proteomic and other applications (EMSL and other laboratories);

- The Mouse House at ORNL;
- The RDP at Michigan State;
- National Centers for High Performance Computing;
- Electron Microscopes;
- Laser-based Imaging Facilities;
- Neutron Facilities at HFIR, LANSCE, and at SNS in future; and

· Technology development for genomics and proteomics, sequencing, gene expression measurement and comparison, etc. in several locations.

In several instances, the impact of existing or expanded facilities could be significantly increased by providing for more efficient access by the wider scientific community. Where appropriate, such improvements in access should become a high priority proximal goal for GTL.

B. New Facilities and Resource Centers:

The challenge of increasing the scale of data acquisition and experimentation and of developing technologies for working at the scale needed for GTL requires that we establish several research facilities/centers. Several of these could immediately be fielded in pilot facilities to test them against real biological problems, train a community of scientists in their use, and to understand the economies of scale and cost-effectiveness, and define the scope, and technical and support requirements of these kinds of facilities. It is important to determine definitively the advantages of a centralized capability in many cases it is clear at the outset, for others it may not be. We believe that the completion of the list, and their prioritization and integration should be addressed immediately by BERAC panels and workshops in the broad scientific community. The facilities or resources listed immediately below are directly coupled to GTL goals and are therefore likely to be high priority.

The list is intended to be illustrative of areas for development, not inclusive.

· Facilities for the Analysis of Multiprotein Molecular Machines that develop and employ technologies for identifying and understanding interactions between cellular proteins and their functions. The technologies required include: Large-scale and high-through-put protein expression and production; sample separation and preparation technology development; sample preparation; instrumentation development; mass spectrometry analyses; and integrated bioinformatics and computation.

· Facilities for Mapping and Modeling Gene Regulatory Networks that initially pilot and subsequently scale up regulatory network discovery and mapping. This involves the development and integration of whole-cell proteomics capabilities, large-scale gene and protein chip analyses,comparative genomics including multi-species large insert libraries, new methods for analyzing cis-regulatory elements in the genome, gene regulatory network bioassays, and an integrated computational and bioinformatics program.

· Facilities for the Analysis of Microbial Growth and Interaction that include chemostats and fermentor farms to study the growth and dynamics of microbial systems in pure and mixed cultures under a variety of conditions. As studies progress these facilities will develop Lab-Bench-scale pilots for investigating various energy production, biomass conversion and carbon sequestration scenarios. Technologies will include: microbial imaging capabilities such as atomic force microscopy and related imaging technologies; environmental scanning electron microscopy to image live-hydrated microbes and related high resolution imaging technologies as well as microchemistry capabilities such as ion microprobe type analyses, focused ion beam, secondary ion mass spectroscopy for high resolution chemical mapping of intra- and extra-cellular enzyme complexes as well as cell wall components of microbes, electron microscopy linked to electron energy loss spectroscopy to facilitate micro-chemical analyses, etc.

· Combinatorial Chemistry Facilities for Small-molecule based Functional Genomics that integrate the design and synthesis of novel small molecule libraries, development of novel high-throughput biological assays, and creation of inventive strategies for identifying gene function. The rapid development of specific inhibitors of protein function, for example, can be powerful reagents in the dissection of protein function.

· Molecular Imaging Facilities to develop new labeling chemistries and imaging capabilities leading to a high throughput capabilities in areas that include: cryo-electron microscopy, soft X-ray microscopy, small-angle X-ray and neutron scattering capabilities, and single molecule detection methods that allow the imaging of several molecules at the same time and that can be used to characterize the functional dynamics of proteins, including their sub-cellular location.

Production Proteomics Facilities that: produce milligram quantities of thousands of proteins for use in function studies, assays and structural analyses; perform high-throughput, global, ultra-sensitive and quantitative measurements of RNA and protein expression; and provide informatics and computational tools to manage, analyze and provide access to the information produced by the PPF and to ensure integration of the PPF data and knowledge base into the systems biology enterprise.

· Mouse Facility The advent of genomic sequence information and revolutionary genome based techniques has given a new importance to studies of the functions of biological systems in- vivo in mice.

Facilities and Resource Center needs include:

1. Advanced technologies for manipulating mouse genes, such as

    a. New Lenti-virus vectors that appear to simplify the process of making transgenic animals;

    b. Ability to temporally manipulate specific gene expression; and

    c. Imaging specific gene products in real time in live animals.

2. User facilities for systematic production of

    a. Transgenic and gene replacement animals;

    b. Mutagenesis, for example the new ethylnitrosourea (ENU) mutagenesis projects that can mutagenize any and all genes in vivo, examine the phenotypic effects by high-throughput screening, and rapidly map the causative gene(s). This later can only be done in large central facilities.

Computational Capabilities -- Linking the computational capabilities of the national labs to the day-to-day operation of Genomes to Life research facilities is one key to making these facilities truly powerful and unique. The computational biology roadmap, jointly developed by ASCR and BER, projects a computational biology enterprise which includes: the development and use of new computing and information technologies for handling, storage and retrieval of data and knowledge, and an aggressive program of focused mathematics and computer science research focused on methods for modeling complex biological systems and comparing models to experimental data, and an appropriate computational infrastructure for these activities. The amalgamation of these efforts requires additional funding, and mathematical analysis, software and algorithm development dedicated to these biological problems.

Potential Pilot Facilities -- There are several potentially exciting developments in the near term that could provide important enabling technologies for GTL and are worthy of careful consideration. Several of these concepts could immediately be fielded in Pilot facilities.

· Effective facilities for Production of Proteins are very important to provide the materials needed for experimentation. These materials are essential for the understanding of protein complexes and protein-based materials including composites. Carrying these proteins on to crystallization and structure determination is another essential production need for those being studied by macromolecular synchrotron x-ray diffraction and/or neutron scattering. These sorts of capacities need not be directly co-located at the synchrotron facilities as once produced and frozen, samples are readily transported;

· Facilities for High-throughput Proteomics are essential for a systems approach to protein synthesis, modification, distribution, and function. Knowing all possible architectural elements of proteins (a goal of structural genomics) will greatly advance our understanding of single domain structures and predictions. When combined with lower resolution techniques (see below) we can begin to understand function at the intermediate scales.

· Intermediate-scale Imaging Facilities. While crystallography yields atomic resolution structure, Cryo-EM has the advantage of being able to look at many different functional states of the complexes. The recent study of the ribosome is a perfect example of the value of both approaches being pursued in parallel. (A little further out in time is the possibility of doing the same sort of experiments with x-rays that would become available from the so-called fourth generation x-ray synchrotron light sources (and such a source is being planned for construction by DOE, called LCLS, that could be operational in about 5-6 years). There is commonality in the computational analysis and algorithms in these two approaches.)

· Centers for the Analysis of Nano-scale Biological Structures. The burgeoning field of nanoscience and nanotechnology offers a multifaceted opportunity when coupled with the revolution in biology. All of life's processes are based on nano-machines and phenomenology at the nanoscale and are encoded in genomes of all organisms. The connection between genes and molecular machines and materials represents a major opportunity for the BER program. A pilot center with an emphasis on the genomic analysis of biological materials is very timely.

* Large scale DNA sequencing capacity for selected elements of whole-genomes will be an essential ingredient in regulatory network discovery and other GTL goals for years to come. Providing facilities specifically designed to enable diverse researchers to have access to high efficiency large scale sequencing is crucial to maximizing GTL discovery and the impact of the DOE sequencing capacity for the nation.

This list of new facilities needs is certainly not inclusive, and it remains to determine which of these are amenable to effective embodiment into resources and facilities for the community. It is time to begin the evaluation of these and other potential resources that are crucial to the future of biology.

Conclusions and a Call for Action

The time has come, because of the genomic revolution, to step forward and assess the new needs of the biological sciences. We have within our reach the ability to bring together the biological, physical, and computing sciences to enable a fundamental understanding of life and the resulting benefits to the nation and the world. BER has played a catalytic and instrumental role in bringing about the genomic revolution and is now poised to make equally seminal contributions to this next, transforming phase. The vision of the BER Genomes to Life Program is designed to build on major accomplishments of the past decade and move from this vision to reality. The next steps are the real payoff to the nation and to the DOE mission. Since the science has changed so profoundly, new facilities and new scientific resources are clearly needed. The DOE should now create unique, high-technology and high-throughput research facilities to translate the new biology, embodied in the goals of the Genomes to Life program, into a reality for the nation. The Genome to Life Program provides both the rationale and nucleus of a broader program to bring the benefits of the genome revolution into reality. We argue here for a new plan that will provide the needed facilities and scientific infrastructure, and call for an immediate planning process including action by BERAC and workshops for evaluation and prioritization of needs.   The implementation of this plan could have a transforming effect on the biological sciences of the next two decades.